

A general method for identifying node spreading influence via the adjacent matrix and spreading rate

Jian-Hong Lin,¹ Jian-Guo Liu,^{1, a)} and Qiang Guo¹

Research Center of Complex Systems Science, University of Shanghai for Science and Technology, Shanghai 200093, P. R. China

(Dated: 27 August 2014)

With great theoretical and practical significance, identifying the node spreading influence of complex network is one of the most promising domains. So far, various topology-based centrality measures have been proposed to identify the node spreading influence in a network. However, the node spreading influence is a result of the interplay between the network topology structure and spreading dynamics. In this paper, we build up the systematic method by combining the network structure and spreading dynamics to identify the node spreading influence. By combining the adjacent matrix A and spreading parameter β , we theoretical give the node spreading influence with the eigenvector of the largest eigenvalue. Comparing with the Susceptible-Infected-Recovered (SIR) model epidemic results for four real networks, our method could identify the node spreading influence more accurately than the ones generated by the degree, K-shell and eigenvector centrality. This work may provide a systematic method for identifying node spreading influence.

PACS numbers: 89.20.Hh, 89.75.Hc, 05.70.Ln

I. INTRODUCTION

Spreading is a widespread process in nature, which describes many important activities in society¹⁻⁴, such as the virus spreading⁵, reaction diffusion process^{6,36}, pandemics⁸, cascading failures⁹ and so on. The knowledge of the spreading pathways through the network of interactions is important for developing effective methods to either hinder the disease spreading, or accelerate the information dissemination spreading. So far, there are a lot of works focusing on identifying the node spreading influence in a network¹⁰⁻¹⁸. Related classical centrality methods include the degree as the number of the node's neighbors, eigenvector centrality¹⁹ as the eigenvector of the largest eigenvalue of the adjacent matrix, K-shell centrality¹ as an effective algorithms based on node location that outperform the classical centrality methods the closeness centrality²⁰ as the reciprocal of the sum of the geodesic distances to all other nodes, betweenness centrality^{21,22} as the number of shortest paths through a certain node. Lately, a lot of works tried to improve the classical methods and proposed effective methods for identifying node spreading influence. For example, Sabidussi²⁰ and Chen *et al*²⁴⁻²⁷ focused on directly improving the basic centrality measures including degree, closeness and betweenness. Liu and Zeng^{13,14} tried to improve the K-shell method by removing the degeneracy of the method. Poulin²⁸ focused to cut down the computational complexity of the eigenvector. Moreover, the concept of path diversity is used to improve the ranking of spreaders²⁹. Liu and Ren^{16,30} also designed in directed networks to identify the influential spreaders such as LeaderRank, which is shown to outperform the well-

known PageRank method in both effectiveness and robustness.

The above classic and improved centrality methods are based on the network topology structure. However, the node spreading influence is determined not only by the network structure but also by the spreading dynamics³¹⁻³⁶. The study of spreading dynamics is a promising domains that is finding more and more applications in a wide range of areas and it also can help us to understand the unfold of dynamical processes in complex networks³⁷. Therefore it is necessary to build up the systematic method to identify the node spreading influence by combining the network structure and spreading dynamics. In this paper, we design a structure spreading dynamics (SSD) method for identifying node spreading influence. Since the adjacent matrix can reflect the network structure, we build up a differential equation by the network adjacent matrix and the spreading process. Then the node spreading influence under different time step t , spreading rate β and recovering rate μ can be identified by function of adjacent matrix A . To evaluate the performance of the SSD method, the Kendall's tau τ is introduced to measure the correlation between the ranking list from different centralities and the ranking list from the true spreading influence. The results show that the SSD method can identify the node spreading influence centrality methods. This work provides a systematic method for ranking the node spreading influence.

II. METHOD

In this section we will introduce some basic connect from graph theory which will be used in the rest of paper.

Normally, An undirect network $G = (N, E)$ with N nodes and E edges could be described by an adjacent matrix $A = \{a_{ij}\}$ where $a_{ij} = 1$ if node i is connected

^{a)}liujg004@ustc.edu.cn



FIG. 1. An example network consisted 3 nodes and 2 edges. Node 1 is an initial infected node. It would infect its neighbour node 2 with probability β and recover with probability μ at time step 1.

by node j , and $a_{ij} = 0$ otherwise. For an undirect network, A is binary and symmetric with zeros along the main diagonal. Therefore, the eigenvalues of A will be real. We label the eigenvalues of A in descending order: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$. Since A is a symmetric and real-valued matrix, $A = Q\Lambda Q^T$, where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$, $Q = [\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n]$ and \mathbf{q}_i is the eigenvector of eigenvalue of λ_i .

Implementing the SIR¹ spreading process for one network, in the SIR model, There are three compartments: (i) Susceptible individuals represent the individuals (not yet infected) who are easy to be infected; (ii) Infected individuals represent individuals who have been infected and are able to spread the disease to susceptible individuals; (iii) Recovered individuals represent individuals who have been recovered and will never be infected again. In each time step, we denote that all nodes are initially susceptible except only one infectious node. The infected nodes will infect their susceptible neighbors with the spreading rate β , and infected nodes would recover with recovering rate μ in the next time step. The number of infections generated by the initially-infected node is denoted as its spreading influence. For each initial node, the node spreading influence is obtained by averaging over 100 independent runs and 10 time steps in Fig. 2-3.

We now introduce the structure spreading dynamics (SSD) method. We build up a systematic method by differential equation by combining the adjacent matrix A spreading process. The basic idea is that an infected node would infect its neighbours with spreading rate β and recover or remove with the recovering rate μ . We denoted $x_i(t)$ is the state of node i at time step t . $\mathbf{x}(0)$ is the initial state of a network. If $\mathbf{x}_i(0) = 1$ and $\mathbf{x}_{j \neq i}(0) = 0$, node i is initial infected node. Therefore, $\mathbf{x}(t) - \mathbf{x}(t-1)$ is the probability of the nodes to be infected at time step t . We can approximate by the linearization

$$\mathbf{x}(t) - \mathbf{x}(t-1) = \beta A[\beta A + (1 - \mu)I]^{t-1} \mathbf{x}(0), \quad (1)$$

where β is the spreading rate, μ is the recovering rate, A is the network adjacent matrix, I is a $N \times N$ unit matrix and $\mathbf{x}(0)$ is the initial state of network. As shown in Fig. 1, node 1 is an initial infected node. Therefore, $\mathbf{x}(0) = [1, 0, 0]^T$ and the probability of the nodes to be infected at time step 1 would be $\mathbf{x}(1) - \mathbf{x}(0) = \beta A \mathbf{x}(0) = [0, \beta, 0]^T$. The total probability $\mathbf{x}(t) - \mathbf{x}(0)$ of the nodes

to be infected at time step t would be

$$\begin{aligned} \mathbf{x}(t) - \mathbf{x}(0) &= \sum_{k=1}^t [\mathbf{x}(k) - \mathbf{x}(k-1)] \\ &= \sum_{k=0}^{t-1} \beta A[\beta A + (1 - \mu)I]^k \mathbf{x}(0). \end{aligned} \quad (2)$$

The node spreading influence of node i , $\mathbf{S}_i(t)$, could be approximate calculated by the following way

$$\mathbf{S}_i(t) = \left\{ \sum_{k=0}^{t-1} \beta A[\beta A + (1 - \mu)I]^k \right\}^T \mathbf{l}_i, \quad (3)$$

where \mathbf{l} is a $N \times 1$ matrix whose components are 1. When recovering rate $\mu = 0$ and $\mu = 1$, $\mathbf{S}_i(t)$ is the spreading influence of node i for SI and standard SIR model at time step t respectively.

The spreading influence of node i , $\mathbf{S}_i(t)$, can be written in the following way by decomposing the adjacent matrix A ,

$$\mathbf{S}_i(t) = m_1 \mathbf{q}_{1i} \sum_{j=1}^n \mathbf{q}_{1j} + \sum_{k=2}^n m_k \mathbf{q}_{ki} \sum_{j=1}^n \mathbf{q}_{1j}, \quad (4)$$

where $m_k = (\mu - \beta \lambda_1) \{ \beta \lambda_k [1 - (\beta \lambda_k + 1 - \mu)] \}^{-1}$. Let $\varphi_i(t) = (m_1 \sum_{j=1}^n \mathbf{q}_{1j})^{-1} \mathbf{S}_i(t)$, Then

$$\varphi_i(t) = \mathbf{q}_{1i} + (m_1 \sum_{j=1}^n \mathbf{q}_{1j})^{-1} \sum_{k=2}^n m_k \mathbf{q}_{ki} \sum_{j=1}^n \mathbf{q}_{1j}, \quad (5)$$

The ranking list generated by $\varphi(t)$ is the same as $\mathbf{S}(t)$. Since $\lambda_1 > \lambda_k$, for $2 \leq k \leq n$, as $\beta \rightarrow 1$ and $t \rightarrow \infty$ we can find that $\varphi(t) \rightarrow \mathbf{q}_1$. By the Perron-Frobenius Theorem³⁸ $\mathbf{q}_1 > 0$. Thus when $\beta \rightarrow 1$ and $t \rightarrow \infty$, the ranking list generated by SSD method is the same as the one generated eigenvector centrality.

III. EXPERIMENT RESULTS

A. Data description

To check the performance of the SSD method, two real networks are introduced in this paper including the Email³⁹ and Protein networks. The Email network of University Rovira i Virgili (URV) of Spain contains faculty, researchers, technicians, managers, administrators, and graduate students. The Protein network is a protein-protein interaction network in budding yeast.

The statistical properties of two real networks are shown in Table I, including the number of nodes N , edges E , the average degree $\langle k \rangle$ and the largest eigenvalue λ_{max} .

TABLE I. Basic statistical features of Email and Protein networks, including the number of nodes N , edges E , the average degree $\langle k \rangle$ and the largest eigenvalue λ_{max} .

Network	N	E	$\langle k \rangle$	λ_{max}
Email	1133	5451	9.60	20.75
Protein	2284	6646	5.82	19.04

B. Measurement

To evaluate the performance of the SSD method, the Kendall's tau τ is introduced to measure the correlation of the node spreading influence with SSD method, degree, K-shell and eigenvector centrality. The Kendall's tau τ is used to measure the correlation between two ranking lists. The Kendall's tau τ value is between $[-1, 1]$, and the increasing values imply the method can identify the node spreading influence more accurately. The Kendall's tau τ is defined as

$$\tau = \frac{2}{N(N-1)} \sum_{i < j} \text{sgn}[(y_i - y_j)(z_i - z_j)], \quad (6)$$

where N is the number of nodes of a network, $y(i)$ is the node spreading influence of node i , $z(i)$ are the values generated by the SSD method, degree, K-shell and eigenvector centrality and $\text{sgn}(x)$ is a piecewise function, when $x > 0$, $\text{sgn}(x) = +1$; $x < 0$, $\text{sgn}(x) = -1$; when $x = 0$, $\text{sgn}(x) = 0$.

C. Numerical results

In this section we check the performance of the SSD method by the Kendall's tau τ . As shown in Fig. 2, the Kendall's tau values τ of the SSD method is between 0.66 and 0.93, which indicates that the ranking list generated by the SSD method are highly identical to the ranking list by the SIR spreading process. The comparisons between the SIR model and the SSD method show that the nodes with influential neighbors will have larger spreading influence. Comparing with degree, K-shell and eigenvector centrality, the Kendall's tau τ of the SSD method would be much better than the ones generated by other methods, which indicates that the SSD method can identify the node spreading influence more accurately than degree, K-shell and eigenvector.

Figure 3 reports the improved ratio in the Kendall's tau τ when applying the SSD method compare with degree, K-shell and eigenvector eigenvector. The improved ratio is defined as

$$\eta = \frac{\tau^S - \tau^0}{\tau^0}, \quad (7)$$

where τ^S is the Kendall's tau of the SSD method, τ^0 is the Kendall's tau of degree, K-shell and eigenvector respectively. Clearly, $\eta > 0$ indicates an advantage of the

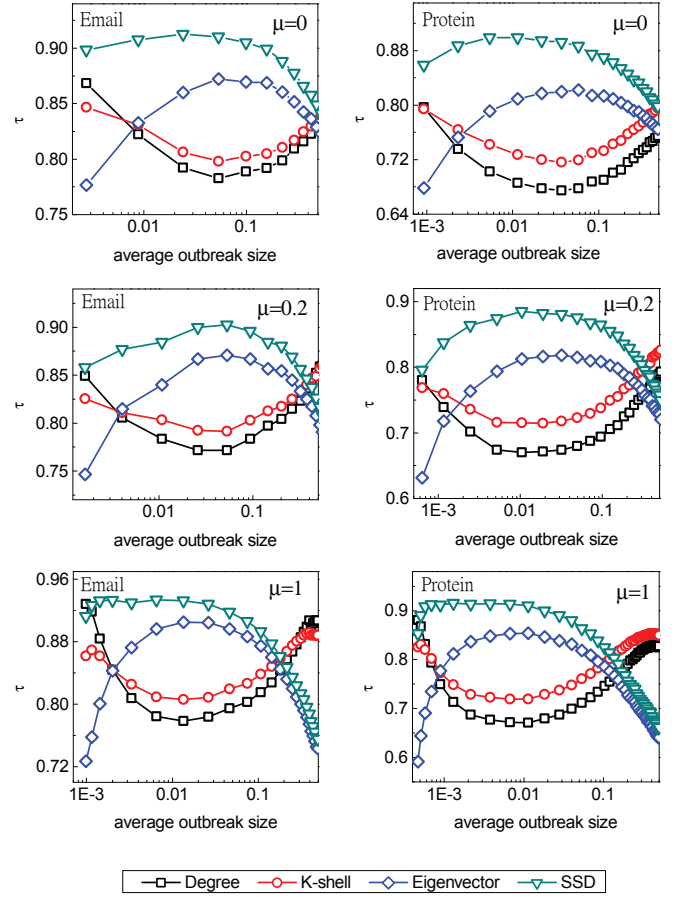


FIG. 2. (Color online) The Kendall's tau values τ obtained by comparing the ranking list generated by the SIR spreading process and the ranking lists generated by the degree (squares), K-shell (circles), eigenvector (diamonds) and SSD method (triangles) with recovering rate μ 0, 0.2, 1 respectively. The average outbreak size (horizontal axis) being controlled by the spreading rate β is the average number of the infected nodes when choosing the initial node of the network. From which one can find that the SSD method could identify the node spreading influence more accurately than other methods. The results are averaged over 100 independent runs with different spreading rate β when the average outbreak size reach 50% of the network.

SSD method. The improved ratio in τ for degree, K-shell and eigenvector with different spreading rate β and recovering rate μ on two real networks are shown in Fig. 4. From which one can find that the ranking accuracy has been remarkably improved by the SSD method in different methods. The largest improved ratio η for degree, K-shell and eigenvector could reach 35.9%, 27.0% and 44.1% respectively.

However we can find that the Kendall's tau τ decreases with the increase of the spreading rate β when the recovering rate $\mu = 1$ for Email and Protein network in Fig. 2 and the improved ratio is even lower than 0 for large spreading rate β , which indicates the SSD method fails to identify the node spreading influence with large spreading

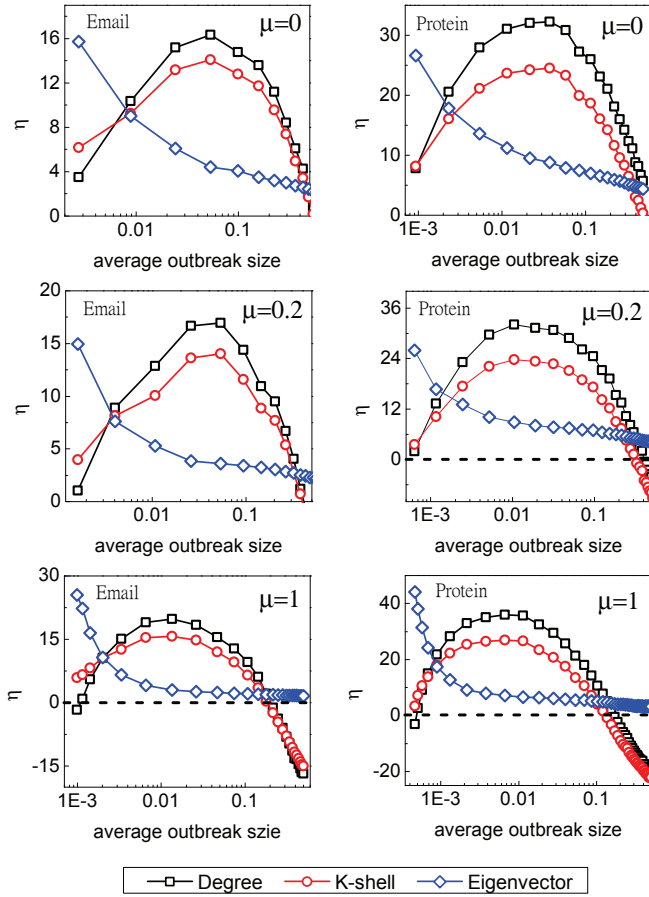


FIG. 3. (Color online) The vertical axis η is the improved ratio η for degree, K-shell and eigenvector centrality with different spreading rate β on two real networks. From which one can find that the improved ratio $\eta > 0$ indicates the Kendall's tau for SSD method is higher than other Kendall's tau generated by other methods. The results are averaged over 100 independent runs with different spreading rate β when the average outbreak size reach 50% of the network.

rate β . Because SSD method is an approximate method for calculating the node spreading influence and there are two disadvantages in SSD method. Firstly, it does not consider the node state at time step $t - 1$ when calculating the probability of the nodes to be infected at time step t by equation (3). Secondly, the SSD method calculates the probability of a node which has two infected nodes to be infected by linear method instead of non-linear method. For example, according to equation (3) if a susceptible node i has two infected neighbour nodes at time step $t - 1$, the probability of node i to be infected at time step t is 2β instead of $1 - (1 - \beta)^2$.

We can find that the curve of SSD method has the same trend with eigenvector centrality. Especially the Kendall's tau τ of the SSD method is the same as eigenvector method with large spreading rate β . Figure 4 reports the correlation between the SSD method and the eigenvector centrality with different spreading rate β and

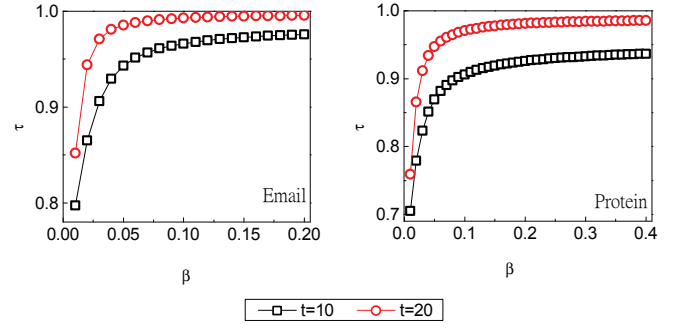


FIG. 4. (Color online) The Kendall's tau values τ obtained by comparing the ranking list generated by the SSD method and the ranking lists generated by eigenvector centrality when the recovering rate $\mu = 0.2$ and time step is 10 (squares) and 20 (circles) for Email and Protein network. From which one can find that the Kendall's tau τ of the SSD method and eigenvector centrality is almost equal to 1 when the spreading rate β and the time step is large, which indicates the ranking list generated by the SSD method have the same ranking list by the eigenvector centrality for Email and Protein network.

time step t when the recovering rate $\mu = 0.2$. From which one can find that the Kendall's tau τ of the ranking list generated by SSD method and eigenvector method increases with the spreading rate β , which indicates the ranking list generated by the SSD method is the same as the one generated by eigenvector method with large spreading rate β and time step t which is proved in the section 3.

IV. CONCLUSION

In this paper, we propose a general framework for identifying the node spreading influence by combining the network structure and the spreading dynamics. By theoretical analyzing the spreading differential equation, one can get that the total number of the infected node for one target node is determined by the adjacent matrix A , spreading parameter β and initial state of the target node. Therefore, we propose a structure spreading dynamics (SSD) method for ranking the node spreading influence. The simulation results for two real networks show that the Kendall's tau τ of the SSD method is between 0.66 and 0.93, which indicates that the ranking list generated by the SSD method is highly identical to the ranking list by the SIR spreading process. Comparing with the degree, K-shell and eigenvector centrality, the largest improved ratio η could reach 35.9%, 27.0% and 44.1% respectively. Furthermore we can find that the ranking list generated by the SSD method is almost the same as the one generated by eigenvector centrality with large spreading rate β and time step t as we analyze.

However, the Kendall's tau τ of the SSD method decreases with the increase the spreading rate β when the recovering rate $\mu = 1$ in Email and Protein network,

which indicates the SSD method could not identify the node spreading influence very well for large spreading rate β . Because SSD method is an approximate method and there are two disadvantages in this method. Firstly, it does not consider the node state at time step $t-1$ when calculating the probability of the nodes to be infected at time step t by equation (3). Secondly, the SSD method calculates the probability of a node which has two infected neighbour nodes to be infected by linear method instead of non-linear method. The solving of the above problems can help us to improve the accuracy of the SSD method for identifying the node spreading influence and study the multiple-nodes spreading process.

ACKNOWLEDGMENTS

The authors wish to thank Dr. Tao Zhou for discussion. This work is supported by the National Natural Science Foundation of China (Nos. 71171136), the Shanghai Leading Academic Discipline Project of China (No. XTKX2012), MOE Project of Humanities and Social Science (No. 13YJA630023), the Foundation of Shanghai Research Institute of Publishing and Media (No. SAYB1407).

- ¹M. Kitsak, L.K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H.E. Stanley, H.A. Makse, *Nat. Phys.* 6 (2010) 888-893.
- ²K. Klemm, M. Serrano, V. Eguiluz, M. Miguel, *Sci. Rep.* 2 (2012) 292.
- ³C. Castellano, R. Pastor-Satorras, *Sci. Rep.* 2 (2012) 371.
- ⁴T. Zhou, J.-G. Liu, W.-J. Bai, G. Chen, B.-H. Wang, *Phys. Rev. E* 74 (2006) 056109.
- ⁵J.O. Kephart, G.B. Sorkin, D.M. Chess, S.R. White, *Sci. Am.* 277 (1997) 56-61.
- ⁶V. Colizza, R. Pastor-Satorras, A. Vespignani, *Nat. Phys.* 3 (2007) 276-282.
- ⁷J.-G. Liu, Z.-X. Wu, F. Wang, *Int. J. Mod. Phys. C* 18 (2007) 1087-1094.
- ⁸R. Pastor-Satorras, A. Vespignani, *Phys. Rev. Lett.* 87 (2001) 258701.
- ⁹C. Castellano, R. Pastor-Satorras, *Sci. Rep.* 2 (2012) 371.
- ¹⁰G. Ghoshal, A.L. Barabási, *Nat. Commun.* 2 (2011) 394.
- ¹¹J. Borge-Holthoefer, Y. Moreno, *Phys. Rev. E* 85 (2012) 026116.
- ¹²J. Borge-Holthoefer, A. Rivero, Y. Moreno, *Phys. Rev. E* 85 (2012) 066123.
- ¹³J.-G. Liu, Z.-M. Ren, Q. Guo, *Physica A* 392 (2013) 4154-4159.
- ¹⁴A. Zeng, C.-J. Zhang, *phys. lett. A* 377 (2013) 1031-1035.
- ¹⁵Z.-M. Ren, F. Shao, J.-G. Liu, Q. Guo, B.-H. Wang, *Acta Phys. Sin.* 62 (2013) 128901.
- ¹⁶Z.-M. Ren, A. Zeng, D.-B. Chen, H. Liao, J.-G. Liu, *Europhys. Lett.* 106 (2014) 48005.
- ¹⁷X.-L. Ren, L.-Y. Lü, *Chinese Sci. Bull.* 59 (2014) 1175-1197.
- ¹⁸C. Orsini, E. Gregori, L. Lenzini, D.Krioukov, *arXiv: 1301.5938v1* (2013).
- ¹⁹S.P. Borgatti, *Soc. Netw.* 27 (2005) 55-71.
- ²⁰G. Sabidussi, *Psychometrika* 31 (1966) 581-603.
- ²¹L.C. Freeman, *Sociometry* 40 (1977) 35-41.
- ²²L.C. Freeman, *Social Netw.* 1 (1979) 215-239.
- ²³J. Ugander, L. Backstrom, C. Marlow, J. Kleinberg, *Sci. USA* 109 (2012) 5962-5966.
- ²⁴D.-B. Chen, L.-Y. Lü, M.-S. Shang, Y.-C Zhang, T. Zhou, *Physica A* 391 (2012) 1777-1787.
- ²⁵D.-B. Chen, H. Gao, L.-Y. Lü, T. Zhou, *PLOS ONE* 8 (2013) e77455.
- ²⁶C. Dangalchev, *Physica A* 365(2006) 556-564.
- ²⁷J. Zhang, X.-K. Xu, K. Zhang, M. Small, *Chaos* 21 (2011) 016107.
- ²⁸R. Poulin, M. C. Boily, B. R. Mâsse, *Social Netw.* 22 (2000) 187-220.
- ²⁹D.-B. Chen, X. R, A. Zeng, Y.-C. Zhang, *Europhys. Lett.* 104 (2013) 68006.
- ³⁰L.-Y. Lü, Y.-C Zhang, T. Zhou, *PLoS ONE* 6 (2011) e21202.
- ³¹J.-G. Liu, Z.-M. Ren, Q. Guo, B.-H. Wang, *Acta Phys. Sin.* 62 (2013) 178901.
- ³²J. Borge-Holthoefer, A. Rivero, Y. Moreno, *Phys. Rev. E* 85(2012) 066123.
- ³³J. Borge-Holthoefer, Y. Moreno, *Phys. Rev. E* 85(2012) 026116.
- ³⁴K. Klemm, M. Á. Serrano, V. M. Eguíluz, M. San Miguel, *Sci. Rep.* 2 (2012) 292.
- ³⁵S. Aral, D. Walker, *Science* 68 (2012) 337-341.
- ³⁶J.-G. Liu, Z.-X. Wu, F. Wang, *Int. J. Mod. Phys. C* 18(2007), 1087-1094.
- ³⁷R. Pastor, C. C. Satorras, P. Van Mieghem, A. Vespignani, *Rev. Mod. Phys.* submitted 2014.
- ³⁸R. A. Horn, C. R. Johnson, *Matrix analysis* (Cambridge University Press, Cambridge) 1985.
- ³⁹R. Guimera, A. Diaz-Guilera, F. Giralt, A. Arenas, 68 (2003) 065103.
- ⁴⁰S.N. Dorogovtsev, A.V. Goltsev, J.F.F. Mendes, *Phys. Rev. Lett.* 96 (2006) 040601.
- ⁴¹S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, E. Shir, *Proc. Natl. Acad. Sci. USA* 104 (2007) 11150-11154.